An Innovative Hybrid Soft Consensus Framework Leveraging Generative Large Language Models and Human Expertise

F.J. Cabrerizo
University of Granada
cabrerizo@decsai.ugr.es

J. Bernabé-Moreno
IBM Research Europe
Juan.Bernabe-Moreno@ibm.com

E. Herrera-Viedma
University of Granada
viedma@decsai.ugr.es

I.J. Pérez
University of Granada
ijperez@decsai.ugr.es

Abstract

Traditional Group Decision Making models have often used artificial intelligence tools, such as fuzzy logic, to act as moderators among multiple human experts tasked with selecting the best option from a set of alternatives. These models facilitated consensus by interpreting and integrating diverse expert opinions. However, the advent of generative Artificial Intelligence models introduces a new challenge: how to integrate these advanced models as active participants in the decision group, alongside human experts. This paper proposes a new hybrid consensus framework that integrates generative large language models as key contributors to the negotiation process. By taking advantage of the unique strengths of both human expertise and AI-driven insights, our framework aims to enhance the robustness and efficiency of group decision making. We explore methods for effectively integrating generative models, addressing potential biases, and ensuring coherent collaboration between human and AI participants. This approach not only enriches the decision making process, but also sets a precedent for future collaborative systems combining human knowledge and artificial intelligence.

Keywords: Group Decision Making, Soft Consensus, Large Language Models, Fuzzy Logic, Generative AI.

1. Introduction

Group decision making (GDM) is a complex process that involves integrating diverse opinions and expertise to reach a consensus. This process is critical in various fields, including healthcare, finance, and policy-making, where collective judgment is essential for making informed and balanced decisions, (Hirokawa & Poole, 1996). Traditional approaches to group decision-making often struggle with effectively managing differing perspectives and ensuring that all voices are adequately considered, leading to potential conflicts and suboptimal outcomes.

Fuzzy logic is a valuable Artificial Intelligence (AI) tool for supporting GDM by modeling the concept of soft consensus, (Cabrerizo, Moreno, et al., 2010). This approach also allows for the representation of uncertainty and partial truth inherent in human opinions and judgments. Fuzzy logic systems can aggregate these nuanced inputs to facilitate a more flexible and inclusive consensus process, effectively bridging the gap between rigid decision rules and the complexity of human reasoning.

The advent of generative LLMs offers a new dimension for enhancing GDM frameworks. On one hand, these new AI tools can be used as a simple complement to traditional GDM models in order to enrich them with new capabilities such as natural language processing and sentiment analysis that make them able to make deeper processing of the expert's preferences. For example, in Trillo et al. (2024), authors use sentiment analysis and LLMs to analyze the experts' behavior during the debate by measuring the positivity and aggressiveness level of their comments and using it later to adjust the weight of their assessments.

On the other hand, with their advanced natural language understanding and generation capabilities, Large Language Models (LLMs) can act as sophisticated participants in the consensus reaching process. By integrating LLMs as active contributors, it is possible to create a hybrid model where both human expertise and AI-driven insights work together to improve overall decision quality. This integration presents new challenges and opportunities, particularly

URI: https://hdl.handle.net/10125/109060
978-0-9981331-8-8

in ensuring coherent and unbiased communication between human and AI agents.

The primary goal of this contribution is to propose a new hybrid consensus framework that incorporates generative LLMs alongside human experts. This framework aims to leverage the strengths of both parties to achieve more robust and efficient consensus outcomes. By addressing key challenges such as bias mitigation and interpretability, this hybrid model aims to set a new standard for GDM models.

To achieve this goal, we used an open source framework called LangChain, whose tools and APIs simplify the process of building LLM-driven applications. In this way, we can have human opinions and model answers under the same system. Thus, we used LangChain as a common interface for both human experts and LLMs, providing a centralized development environment. In addition, we have taken into account the factors to consider when selecting an LLM, published in Mandvikar (2023). In such a way, we selected the most appropriate ones to perform the different experiments to test and validate the proposed framework over a real-world case of use.

The remainder of this article is organized as follows: Section 2 presents the related works, providing background on fuzzy logic and LLMs reasoning in decision-making contexts. Section 3 details the development of the hybrid consensus model. Section 4 provides an illustrative example of the model in a real-world scenario. Finally, Section 5 discusses conclusions and future work, outlining potential areas for further research and application.

2. Related Works

In this section, a brief summary of different kinds of related approaches is presented. Firstly, the use of fuzzy logic in classical GDM approaches is analyzed. Secondly, other AI tools that can be used as a complement to improve the GDM models as sentiment analysis or LLMs are introduced. Finally, some systems that use just LLMs to accomplish the GDM tasks are presented.

2.1. Fuzzy logic in classical group decision making.

In group GDM problems, the consensus process is an iterative mechanism consisting of multiple rounds of discussion. During these rounds, experts are asked to adjust their preferences based on the guidance provided by the moderator, (Pérez et al., 2020). The moderator plays a central role in this process. Typically, the moderator is a person who does not actively participate in the discussions, but who has extensive knowledge of each expert's preferences and the overall level of agreement throughout the consensus process. The moderator's primary responsibility is to oversee and guide the process toward successful consensus. This includes maximizing the level of agreement and systematically reducing the number of experts whose preferences remain far away from collaborative ones in each iteration.

Fuzzy logic has been a cornerstone of classical GDM models, particularly for its ability to handle uncertainty and approximate reasoning. Traditional decision-making approaches have faced the challenge of integrating diverse and sometimes conflicting expert opinions. Fuzzy logic addresses this by allowing for partial membership in decision sets, allowing for more detailed interpretations of expert input. This approach derives the concept of soft consensus, where agreement is not strictly binary but can vary in degree. Soft consensus is particularly valuable in group settings because it accommodates a range of opinions and helps moderate discussions to converge on a solution that reflects the collective judgment. Therefore, by using fuzzy logic, GDM frameworks can better manage ambiguity and support more flexible, inclusive consensus-building processes, (Cabrerizo, Moreno, et al., 2010; Kacprzyk & Nurmi, 1998).

An essential aspect of applying fuzzy logic to GDM is the use of linguistic labels to represent and analyze expert opinions, (Morente-Molinera et al., 2015). Linguistic labels, such as "most likely", "possible", or "unlikely", provide a more intuitive and human-readable way to express degrees of uncertainty and preference. These labels are transformed into fuzzy sets, which can then be mathematically manipulated to aggregate and compare different opinions. This method allows for a more seamless integration of qualitative judgments into the decision-making process. By using linguistic labels, fuzzy logic systems can capture the subtleties of expert judgment and provide a more natural and accessible means of reaching a high soft consensus level. This approach not only increases the inclusiveness of the GDM process, but also improves the overall quality of the consensus by ensuring that all perspectives are adequately represented and considered.

A fuzzy linguistic preference relation (FLPR) P is defined over a set of alternatives $A = \{a_1, a_2, \ldots, a_n\}$ and is represented by a matrix $P = [p_{ij}]$ of dimension $n \times n$, where each element p_{ij} is a linguistic value indicating the preference of element a_i over element a_j, (Pérez et al., 2011a).

Linguistic values are qualitative terms such as "very preferred", "moderately preferred", "equal," etc. These

terms can be converted into fuzzy numbers for easier mathematical handling. A common approach is to use trapezoidal or triangular fuzzy sets to represent these linguistic terms.

For example, a set of linguistic values S could be:

$$S = \{s_1, s_2, \ldots, s_m\}$$

where each s_k represents a linguistic term.

Suppose we have a linguistic scale S with five terms: $S = \{s_1, s_2, s_3, s_4, s_5\}$, where:

- s_1 = "very low preference"
- s_2 = "low preference"
- s_3 = "equal preference"
- s_4 = "high preference"
- s_5 = "very high preference"

Each term can be represented by a fuzzy number, for example:

- $s_1 \rightarrow (0, 0, 0.1, 0.2)$
- $s_2 \rightarrow (0.1, 0.2, 0.3, 0.4)$
- $s_3 \rightarrow (0.3, 0.4, 0.6, 0.7)$
- $s_4 \rightarrow (0.6, 0.7, 0.8, 0.9)$
- $s_5 \rightarrow (0.8, 0.9, 1, 1)$

The fuzzy linguistic preference relation P could have the following form for a set of three alternatives:

$$P = \begin{pmatrix} - & s_1 & s_3 \\ s_3 & - & s4 \\ s_5 & s2 & - \end{pmatrix}$$

2.2. Human-in-the-loop decision making with automation.

Human-in-the-loop (HITL) decision making with automation is a well-established concept where human judgment is integrated with automated systems to enhance overall decision accuracy and efficiency. Historically, HITL approaches have been pivotal in fields where complex, unpredictable scenarios require human intuition and experience to guide and improve automated processes. This concept originated in early computational models from the 1970s, (Dunn, 1973), where researchers recognized the limitations of fully automated systems in handling ambiguous or nuanced situations. Initially, these approaches primarily focused on AI as an assistive tool rather than an active participant.

Over the decades, the advancement of machine learning and AI has led to more sophisticated implementations of HITL, particularly in areas like medical diagnostics, financial forecasting, and autonomous driving, where human oversight is crucial for safety and ethical considerations. Recent literature, such as Fosch-Villaronga et al. (2021) and Sele and Chugunova (2024), not only has further explored the integration of HITL in modern AI systems, emphasizing enhanced performance through adaptive learning algorithms that learn from human feedback in real-time, but also has concluded that this framework decreases accuracy but increases uptake of automated decision making.

Our research builds upon these foundations but extends them by integrating different LLMs as active decision-making agents within the HITL framework. Unlike previous models, our approach leverages the unique capabilities of LLMs to interpret and generate complex language-based inputs, thereby enhancing the collaborative decision-making process and offering new insights into potential AI-human synergies. Key challenges such as bias mitigation are directly addressed, highlighting the distinctiveness and novelty of our work.

2.3. Sentiment Analysis and LLMs in classical group decision making.

On one hand, some approaches focus on the use of Sentiment Analysis (SA) in GDM processes, (Trillo et al., 2023). In these papers, SA is used to categorize experts' comments during debates as aggressive or non-aggressive, helping to understand their behavior and improve consensus measures. By incorporating linguistic features and machine learning techniques, the proposed method assigns weights to experts' opinions, improving the decision-making process by considering both the content and emotional tone of the debate.

On the other hand, some approaches claim that it is not necessary to develop and train a complete SA algorithm for the task described above. It is sufficient to include a large pre-trained model by using its API, which can help to categorize comments in an efficient way, (Trillo et al., 2024). LLMs offer significant advantages in this context. They are able to process and analyze large amounts of text quickly and accurately, understand nuanced language, and detect subtleties in sentiment that may be missed by simpler models. Furthermore, the integration of LLMs can facilitate real-time sentiment analysis and feedback, helping to

dynamically adjust the decision-making process based on the evolving sentiments and arguments presented during debates, thereby improving the overall efficiency and accuracy of GDM processes.

2.4. Reasoning and debating with LLMs.

Large pre-trained models exhibit different capabilities due to their diverse training data and architectural variations. This has spurred the development of ensembles, (Sagi & Rokach, 2018) in multimodal learning, (Li et al., 2022). The Mixture of Experts technique, a popular ensemble learning method, trains multiple smaller specialized models to increase robustness and overall accuracy, (N. Du et al., 2022). Specific to language models, Self-Consistency, (X. Wang et al., 2023) generates multiple reasoning paths using CoT and selects the most consistent answer as the final output. In (Jiang et al., 2023), the authors propose LLM-Blender, a method for ranking and merging outputs from different models.

Advances in LLM have catalyzed the development of sophisticated prompting and fine-tuning techniques to address reasoning problems, for example, L. Wang et al. (2023). While eliciting reasoning from a single agent is promising, it is fundamentally limited by the lack of diverse insights.

In this research line, several recent approaches have been focused on multi-agent debating frameworks, for example Y. Du et al. (2023), but they are predominantly limited to multiple instances of the same underlying language model.

Unlike these approaches, in Chih-Yao Chen et al. (2023) the authors analyze communication through explanations between different LLM agents and their ability to discuss and persuade each other to improve collective reasoning. In such a way, this approach presents a framework called ReConcile to enhance the reasoning capabilities of LLMs through a round-table conference among diverse LLM agents. Inspired by the ”society of minds” concept, ReConcile fosters collaborative discussions between LLMs over multiple rounds, where agents aim to convince each other to improve their answers. The framework incorporates human-correcting explanations and a confidence-weighted voting mechanism to achieve better consensus. Each discussion round involves generating revised explanations and answers based on previous rounds, taking into account confidence scores, and leveraging demonstrations of convincing human explanations. Furthermore, in multi-model multi-agent systems, estimating each agent's confidence and their ability to defend or improve their opinions become critical components due to individual differences. However, a notable drawback of ReConcile is that the questions addressed are limited to binary yes/no answers, rather than classifying multiple alternatives. This limitation may restrict the applicability of the framework to more complex reasoning tasks that require nuanced or multi-class responses.

By looking at Figure 1, it is possible to get an idea of the main differences between the different proposals that exist in the literature in this research line.

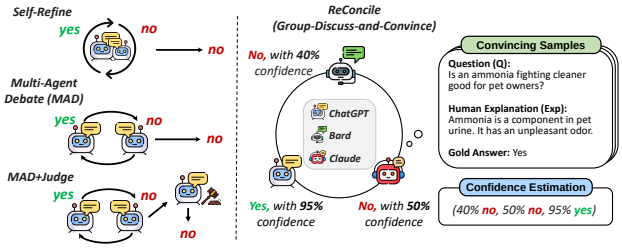

Figure 1. Chih-Yao Chen et al. (2023) The main differences among previously introduced approaches

As a conclusion to this section of related work, it is worth noting that there exist several kinds of AI systems that help people make decisions in different ways:

- Systems that help people to make individual decisions.
- Systems that help to reach consensus in GDM problems
- Systems that use an LLM to help in any way with the decision-making problem
- Systems that integrate different LLMs to debate about the best alternative from different points of view.

However, there are no hybrid models, which take the best of the above systems, integrating the best of both people and LLMs working and reasoning together through the same system.

3. Framework Architecture and process workflow

This section describes the architecture and the workflow of the system.

3.1. Framework architecture

As the primary orchestrator of the proposed architecture, we consider the use of LangChain. LangChain is an open-source, foundational framework

that offers several advantages for developing our hybrid consensus model that integrates LLMs and human experts. In this way, LangChain provides a flexible and scalable infrastructure that simplifies the deployment and orchestration of multiple AI models. Its modular architecture allows seamless integration of different LLMs, facilitating experimentation with different sets of models to determine the most effective combinations. In addition, LangChain's robust support for natural language processing and understanding enables the development of sophisticated communication protocols between AI agents and human participants. This increases the overall coherence and effectiveness of our decision-making process, (Mahadevan, Raman, et al., 2023). By leveraging LangChain advantages, our system can efficiently combine the strengths of both AI and human expertise, ensuring a more robust and dynamic approach to reaching consensus.

3.2. Process workflow

The process workflow is determined by Algorithm 1, which is detailed in the following subsections

3.2.1. Initial Review:

- **Human Experts**: Each human expert reviews a subset of alternatives. They provide initial scores and qualitative feedback based on their expertise in the subject matter but taking into account the different criteria.
- **LLMs**: Using Langchain to join the GDM system, each LLM reviews the same alternatives. The LLMs' assessments include a structured analysis focusing on similar criteria as the human experts.

The structure of the initial scores may depend on each particular problem, but for clarity and consistency of the data, both experts and LLMs should use the same normalized structure of preferences. The most commonly used structures, each one with its own advantages and drawbacks, are presented in Chiclana et al. (1998).

3.2.2. Preliminary Scoring:

- **Combining Insights**: The initial scores and feedback from both human experts and LLMs are compiled. The framework ensures that both sets of scores are normalized to maintain consistency and comparability. This combination depends on the chosen preference representation structure. For this, it will normally be necessary to use an

Algorithm 1 Hybrid Consensus Framework for GDM problems

Input: Set of alternatives $\{Alt_1, Alt_2, ..., Alt_n\}$ and criteria $\{Crit_1, Crit_2, ..., Crit_n\}$
Output: Final selected alternatives $\{Alt_{selected}\}$
Initial Review:
for each alternative Alt_i **do**
    Human Experts: Assign alternative Alt_i to a subset of human experts
    Each expert reviews Alt_i and provides initial scores and qualitative feedback
    LLMs: Use Langchain to review Alt_i using APIs to connect with different pre-trained models
    LLMs provide preliminary assessment taking into account the different criteria
end for
Preliminary Scoring:
Combine scores and feedback from human experts and LLMs
Normalize scores and check consistency
Identify and mitigate biases in LLMs' assessments
Check Consensus level
Consensus Reaching:
repeat
    Deliberation Phase: Conduct virtual meeting with human experts and LLMs
    Experts discuss their views, LLMs provide additional insights
    Negotiation and Re-evaluation: Experts and LLMs are asked to re-evaluate alternatives based on new criteria or perspectives
    Reassess alternatives based on the new insights
    Combine new scores and feedback
    Normalize scores, check consistency and address any new biases
until required consensus level is reached or a predetermined number of iterations is completed
Final Selection:
Synthesize combined recommendations of human experts and LLMs
Rank alternatives based on final scores and qualitative feedback
Final review for coherence and alignment with the decision-making criteria
Output: List of selected alternatives $\{Alt_{selected}\}$

aggregation operator from the OWA family, that suits the particular needs (use of lingustic labels, different weights, etc.).

- **Addressing Biases**: In this case, a human expert point of view is needed to identify and mitigate any biases in the LLMs' assessments. This step is crucial to ensure that the AI's contributions are fair and unbiased.
- **Check Consensus Level**: In order to be able to decide when a satisfactory level of consensus has been reached, allowing a decision to be made in accordance with the assessments of both experts and different LLMs without distancing ourselves too far from any particular opinion, we need to establish the calculation of a soft consensus measure.

  Again, this calculation depends on the chosen preference representation structure, together with different measures of proximity between the individual opinions and the collective opinion. A good example of how such a consensus measure can be computed can be studied in Cabrerizo, Moreno, et al. (2010) and del Moral et al. (2018)

3.2.3. Consensus Reaching: If the consensus level is not high enough, or a predetermined number of iterations has not been completed:

- **Deliberation Phase**: Human experts and LLMs engage in a structured deliberation process. Using the Langchain platform, which shows to the participants the different distance measures computed in the previous step, a virtual meeting is simulated where human experts discuss their views, and LLMs provide additional insights and counterpoints based on their analysis. The discussion is guided by a moderator to ensure all perspectives are considered.
- **Negotiation and Re-evaluation**: Throughout this phase, experts and LLMs are asked to re-evaluate certain alternatives based on new criteria or perspectives discussed during the meeting. The models' ability to generate detailed responses facilitates a dynamic and iterative decision-making process. This re-evaluation includes considering new evidence or re-analyzing data with adjusted parameters.

3.2.4. Final Selection: If the consensus level is high enough, or a predetermined number of iterations has been completed:

- **Synthesis of Recommendations**: The framework synthesizes the combined recommendations of human experts and LLMs. Alternatives that receive high consensus scores from both human and AI participants are prioritized for selection. The synthesis process involves creating a ranked list based on the final scores and qualitative feedback, (Herrera et al., 1995).
- **Ensuring Coherence**: The final list of selected alternatives undergoes a final review to ensure coherence and alignment with the decision-making criteria. This includes ensuring a balanced representation of options and methodologies, as well as checking for redundancy and overlap.

Finally, to illustrate this workflow, A sequence diagram of the proposed system can be observed in Figure 2.

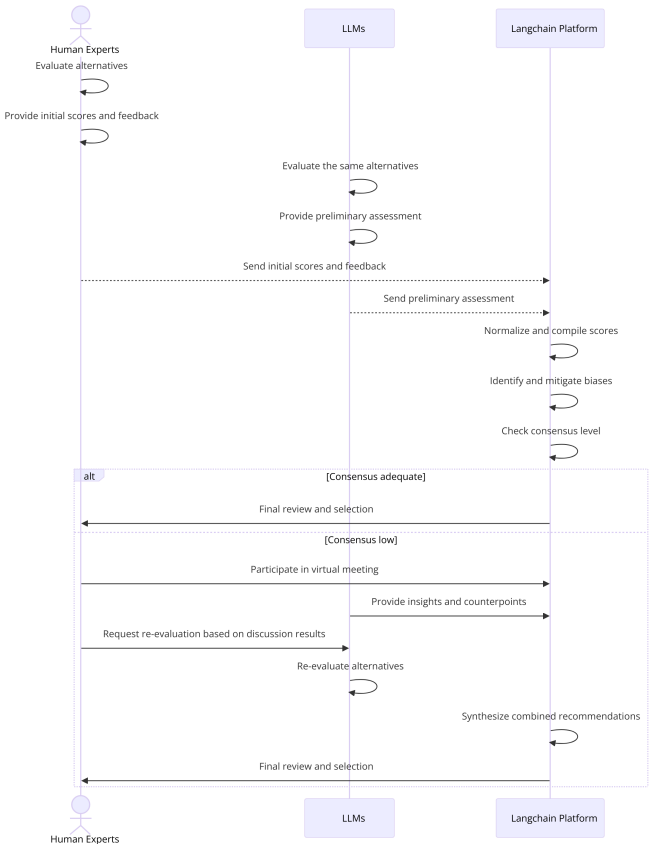

Figure 2. Sequence Diagram

3.3. Outcome

The proposed hybrid consensus framework enhances the robustness and efficiency of the decision-making

process. By leveraging the strengths of both human expertise and AI-driven insights, the GDM system can make more informed and balanced decisions. This approach not only improves the quality and diversity of the selected alternatives but also demonstrates a scalable method for integrating AI into collaborative decision-making systems. The process ensures a transparent and accountable selection procedure, fostering trust and credibility in the decisions made.

4. Real World Scenario: Editorial Selection Process for an Academic Journal

The editorial board of a prestigious academic journal is tasked with selecting the best articles for publication in the upcoming issue. This decision-making process involves different human editors, each with their own expertise and preferences, and several generative (LLMs) integrated via Langchain.

1. **Initial Review**:
    - **Human Editors**: Each human editor reviews a subset of 4 submitted articles. They provide initial scores and qualitative feedback based on their expertise in the field. Specific scoring criteria include relevance to the decision context, novelty of the alternative, methodological rigor, and clarity of presentation. The chosen preference representation structure is the FLPR, using the details presented in Section 2.1. Therefore, each expert should express a single FLPR like this, in addition to a free comment input:

$$P_1 = \begin{pmatrix} - & s_3 & s_1 \\ s_3 & - & s4 \\ s_5 & s2 & - \end{pmatrix}$$

    - **LLMs**: Using Langchain, each LLM reviews the same articles. The models are pre-trained on large academic databases, allowing them to make a preliminary assessment of the articles according to the pre-defined criteria. In the same way, each LLM is asked to format his own report in the same structure as human preferences.
2. **Preliminary Scoring**:
    - **Combining Insights**: Using the LOWA operator, the initial scores of both human editors and LLMs are combined to produce a collective FLPR.
    - **Addressing Biases**: The editor-in-chief is in charge of identifying and mitigating any biases in the LLMs' assessments. This step is crucial to ensure that the AI's contributions are fair and unbiased, (Lin et al., 2024).
    - **Checking Consensus Level:** In this case, we have selected the consensus measure from Pérez et al. (2011b), that is:
        (a) For each pair of experts, a similarity matrix is computed.
        (b) A consensus matrix, CM, is calculated by aggregating all the similarity matrices using the arithmetic mean as the aggregation function
        (c) Once the similarity and consensus matrices have been calculated, the next step involves determining the consensus degrees at three distinct levels. These individual consensus degrees are then combined to derive an overall consensus degree, referred to as the consensus on the relation, CR. This value has to be in the interval [0-1].

      If the consensus measure CR has not attained the stipulated minimum consensus level CL and the number of rounds has not exceeded the predetermined maximum iterations (defined before the commencement of the decision process), it is imperative to revise the expert opinions that are hindering agreement.
3. **Consensus Building**:
    - **Deliberation Phase**: Human editors and LLMs engage in a structured deliberation process. Using the Langchain platform, a virtual meeting is held where human editors discuss their views, and LLMs provide additional insights and counterpoints based on their analysis.
    - **Negotiation and Re-evaluation**: Throughout this phase, editors and LLMs are asked to re-evaluate certain articles based on new criteria or perspectives discussed during the meeting. The models' ability to generate nuanced responses facilitates a dynamic and iterative decision-making process.
4. **Final Selection**:
    - **Synthesis of Recommendations**: The framework synthesizes the combined

recommendations of human editors and LLMs. Articles that receive high consensus scores from both human and AI participants are prioritized for publication. To do so, in order to establish a consensual ranking of the different papers, the selection process defined in Cabrerizo, Heradio, et al. (2010) has been developed.

- **Ensuring Coherence**: The final list of selected articles undergoes a final review from the editor-in-chief to ensure coherence and thematic alignment with the journal's focus and upcoming issue's theme.

A summary of this process for selecting the best articles for publication can be observed in Figure 3.

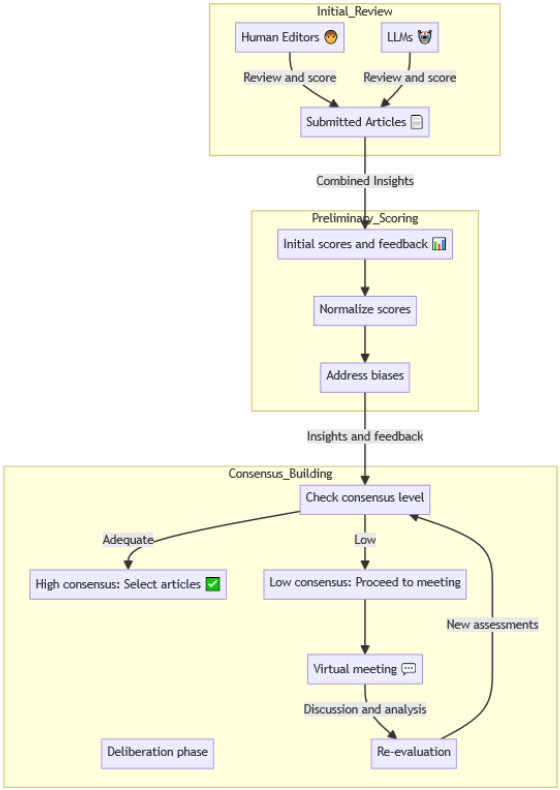

Figure 3. Selecting the best articles for publication process

5. Discussion of results

This section presents the key findings from the experimental validation of our proposed hybrid consensus framework, which integrates Large Language LLMs as active participants alongside human experts in the group decision-making process. The primary aim of these experiments was to assess both the speed of convergence reaching consensus and the overall quality of the involved decisions. To achieve this, we simulated decision-making scenarios where human editors and LLMs collaboratively reviewed and selected the best articles for publication by following the previous section steps. The framework's performance was evaluated based on how quickly consensus was achieved and how well human and AI contributions aligned during the process.

To provide a more comprehensive evaluation of our framework, we compared its performance against a previously published model Pérez et al. (2020), which employed a more traditional approach to consensus building based on gamification techniques to increase the expert´s experience. The results, illustrated in Figure 4, showcase key metrics such as convergence speed and consensus level over several decision-making rounds. This comparison emphasizes the improvements offered by our model, particularly in terms of decision-making efficiency, while also highlighting its superior flexibility and adaptability to dynamic decision-making environments.

We conducted multiple simulations involving different numbers of experts (humans or LLMS) and options to thoroughly examine the model's convergence behavior. In these simulations, the consensus among experts increased progressively with each discussion round, as shown in Figure 4. This indicates that the model facilitates higher levels of agreement among experts as the process advances, underscoring its effectiveness in guiding decision-making toward a more unified outcome.

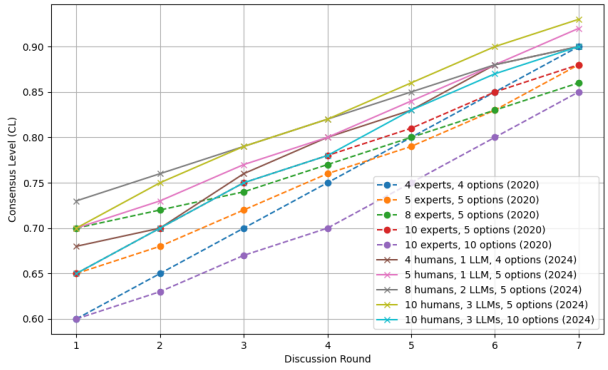

Figure 4. Consensus Level comparative in successive discussion rounds.

6. Conclusions and future work

This study presents a pioneering hybrid consensus framework that incorporates generative LLMs alongside human experts in group decision-making processes. Traditional models have primarily relied on AI tools, such as fuzzy logic, to moderate and integrate human expert opinions. However, the integration of generative AI models as active participants introduces new dynamics and opportunities for enhancing decision-making processes.

The proposed framework demonstrates that combining the unique capabilities of generative LLMs with the nuanced expertise of human participants can lead to more robust and efficient consensus outcomes. By addressing challenges such as potential biases, communication coherence, and the roles of AI and human participants, our approach ensures a balanced and effective negotiation process.

This innovative framework sets a precedent for future research and application in collaborative decision-making systems. It highlights the importance of interdisciplinary approaches that integrate artificial intelligence with human insights, paving the way for more sophisticated and adaptive decision-making environments. Further work is needed to refine integration techniques, improve the interpretability of AI contributions, and ensure seamless interaction between human and AI participants in diverse decision-making scenarios.

Future research will focus on refining the integration of generative LLMs into group decision-making frameworks. Key areas for exploration include developing automatic algorithms to mitigate potential biases introduced by AI models and enhancing the interpretability of their contributions, (Brown, 2024). Additionally, research on the dynamics of human-AI interaction in various decision-making contexts will be crucial, (Echterhoff et al., 2024). This involves understanding how different types of LLMs and human experts can best complement each other and identifying optimal strategies for facilitating seamless and effective communication between them. Finally, Another promising direction is the application of this hybrid consensus framework across diverse practical domains, such as healthcare, finance, and policy-making, to evaluate its versatility and impact.

7. Acknowledgements

This contribution is supported by the grant PID2022-139297OB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU. Moreover, it is part of the project C-ING-165-UGR23, co-funded by the Regional Ministry of University, Research and Innovation and by the European Union under the Andalusia ERDF Programme 2021-2027.

References

Brown, N. B. (2024). Enhancing trust in llms: Algorithms for comparing and interpreting llms. *arXiv preprint arXiv:2406.01943*.

Cabrerizo, F. J., Heradio, R., Javier Perez, I., Herrera Viedma, E., et al. (2010). A selection process based on additive consistency to deal with incomplete fuzzy linguistic information.

Cabrerizo, F. J., Moreno, J. M., Perez, I. J., & Herrera-Viedma, E. (2010). Analyzing consensus approaches in fuzzy group decision making: Advantages and drawbacks. *Soft Computing*, *14*, 451–463.

Chiclana, F., Herrera, F., & Herrera-Viedma, E. (1998). Integrating three representation models in fuzzy multipurpose decision making based on fuzzy preference relations. *Fuzzy sets and Systems*, *97*(1), 33–48.

Chih-Yao Chen, J., Saha, S., & Bansal, M. (2023). Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv e-prints*, arXiv–2309.

del Moral, M. J., Chiclana, F., Tapia, J. M., & Herrera-Viedma, E. (2018). A comparative study on consensus measures in group decision making. *International Journal of Intelligent Systems*, *33*(8), 1624–1638.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. (2022). Glam: Efficient scaling of language models with mixture-of-experts. *International Conference on Machine Learning*, 5547–5569. https : / / proceedings.mlr.press/v162/du22c/du22c.pdf

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate. https://arxiv.org/abs/2305.14325

Dunn, R. (1973). Graphics, problem solving and virtual systems. *Proceedings of the June 4-8, 1973, national computer conference and exposition*, 23–30.

Echterhoff, J., Liu, Y., Alessa, A., McAuley, J., & He, Z. (2024). Cognitive bias in high-stakes

decision-making with llms. *arXiv preprint arXiv:2403.00811*.

Fosch-Villaronga, E., Khanna, P., Drukarch, H., & Custers, B. H. (2021). A human in the loop in surgery automation. *Nature Machine Intelligence*, *3*(5), 368–369.

Herrera, F., Herrera-Viedma, E., & Verdegay, J. L. (1995). A sequential selection process in group decision making with a linguistic assessment approach. *Information Sciences*, *85*(4), 223–239.

Hirokawa, R. Y., & Poole, M. S. (1996). *Communication and group decision making*. Sage Publications.

Jiang, D., Ren, X., & Lin, B. Y. (2023). LLM-blender: Ensembling large language models with pairwise ranking and generative fusion. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14165–14178. https://aclanthology.org/2023.acl-long.792

Kacprzyk, J., & Nurmi, H. (1998). Group decision making under fuzziness. In *Fuzzy sets in decision analysis, operations research and statistics* (pp. 103–136). Springer.

Li, S., Du, Y., Tenenbaum, J. B., Torralba, A., & Mordatch, I. (2022). Composing ensembles of pre-trained models via iterative consensus. *International Conference on Learning Representations (ICLR)*. https : / / arxiv . org/abs/2210.11522

Lin, L., Wang, L., Guo, J., & Wong, K.-F. (2024). Investigating bias in llm-based bias detection: Disparities between llms and human perception. *arXiv preprint arXiv:2403.14896*.

Mahadevan, R., Raman, R. C., et al. (2023). Comparative study and framework for automated summariser evaluation: Langchain and hybrid algorithms. *arXiv preprint arXiv:2310.02759*.

Mandvikar, S. (2023). Factors to consider when selecting a large language model: A comparative analysis. *International Journal of Intelligent Automation and Computing*, *6*(3), 37–40.

Morente-Molinera, J. A., Pérez, I. J., Ureña, M. R., & Herrera-Viedma, E. (2015). On multi-granular fuzzy linguistic modeling in group decision making problems: A systematic review and future trends. *Knowledge-Based Systems*, *74*, 49–60.

Pérez, I. J., Cabrerizo, F. J., & Herrera-Viedma, E. (2011a). Group decision making problems in a linguistic and dynamic context. *Expert Systems with Applications*, *38*(3), 1675–1688.

Pérez, I. J., Cabrerizo, F. J., & Herrera-Viedma, E. (2011b). A mobile group decision making model for heterogeneous information and changeable decision contexts. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *19*(supp01), 33–52.

Pérez, I. J., Garcia-Sanchez, P., Cabrerizo, F. J., & Herrera-Viedma, E. (2020). An approach toward a feedback mechanism for consensus reaching processes using gamification to increase the experts' experience. *HICSS*, 1–10.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(4), e1249. https://wires.onlinelibrary.wiley.com/doi/10.1002/widm.1249

Sele, D., & Chugunova, M. (2024). Putting a human in the loop: Increasing uptake, but decreasing accuracy of automated decision-making. *Plos one*, *19*(2), e0298037.

Trillo, J. R., Cabrerizo, F. J., Perez, I. J., Morente-Molinera, J. A., & Herrera-Viedma, E. (2024). A new consensus reaching method for group decision-making based on the large language model gemini for detecting hostility during the discussion process. *IEEE International conference on Evolving and Adaptive Intelligent Systems 2024 (IEEE EAIS 2024)*.

Trillo, J. R., Herrera-Viedma, E., Morente-Molinera, J. A., & Cabrerizo, F. J. (2023). A group decision-making method based on the experts' behavior during the debate. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., & Lim, E.-P. (2023). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2609–2634. https://aclanthology.org/2023.acl-long.147

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=1PL1NIMMrw